УДК 519.85

А.И. Косолап
Украинский государственный химико-технологический университет, г. Днепропетровск
Украина, 49005, г. Днепропетровск, пр. Гагарина, 8

Метод точной квадратичной регуляризации в задачах кластеризации данных

A.I. Kosolap
The Ukrainian State Chemical-Technological University, c. Dnipropetrovsk
Ukraine, 49005, c. Dnipropetrovsk, Gagarina st., 8

Method of an Exact Quadratic Regularization Into Clustering Problem of Data

А.І. Косолап
Український державний хіміко-технологічний університет, м. Дніпропетровськ
Україна, 49005, м. Дніпропетровськ, пр. Гагаріна, 8

Метод точної квадратичної регуляризації в задачах кластерізації даних

В работе рассматривается задача кластеризации данных, в которой множество точек в *n*-мерном пространстве покрывается непересекающимися шарами – кластерами. Эта задача сводится к максимизации нормы вектора на невыпуклом допустимом множестве. Для решения оптимизационной задачи используется метод точной квадратичной регуляризации, который показал преимущество над генетическими и эволюционными методами при решении многочисленных тестовых задач.
Ключевые слова: кластеризация данных, оптимизация,
метод точной квадратичной регуляризации.

In this paper, we consider a problem clustering of data. The set of points cover of spheres in space *n*-dimensional. This problem is reduced to of vector norm maximization on feasible nonconvex set. Then we use a method of an exact quadratic regularization for the solution of an optimizing problem which has shown its superiority over genetic and evolution methods at the solution of numerous test problems.
Key words: clustering problem of data, optimization, method of an exact quadratic regularization.

В роботі розглядається задача кластеризації даних, в якій множина точок у *n*-вимірному просторі покривається кулями, що не перетинаються. Ця задача зводиться до максимізації норми вектору на неопуклій допустимій множині. Для розв’язку оптимізаційної задачі використовується метод точної квадратичної регуляризації, який показав перевагу над генетичними та еволюційними методами при розв’язку багатьох тестових задач.
Ключові слова: кластеризація даних, оптимізація, метод точної квадратичної регуляризації.

Введение

Одной из основных задач искусственного интеллекта и теории распознавания образов является разбиение данных на кластеры [1], [2]. Необходимо объединить данные в группы однотипных объектов – кластеры. Как правило, данные (объекты) представляются точкой в *n*-мерном пространстве, которые необходимо разбить на заданное число подмножеств непересекающимися шарами. Компоненты точек опре-

деляют параметры, которые характеризуют объекты. Расстояние между двумя точками внутри одного кластера всегда меньше расстояния между точками разных кластеров. В качестве расстояния может выбираться различная метрика пространства. Существуют эффективные методы разбиения данных на два кластера. Однако разбиение множества точек на более чем два кластера, представляет собой сложную задачу [1-3]. Это связано с тем, что оптимальное разбиение точек на кластеры сводится к решению много-экстремальной задачи, в которой необходимо найти точку глобального экстремума. В настоящее время для решения таких задач чаще использует генетические или эволюционные методы, которые основаны на случайном поиске и позволяют находить оптимальное решение задач кластеризации только с некоторой вероятностью [4], [5]. Кроме того, эти методы содержат большое количество параметров, от значений которых зависит их эффективность. В работе использован новый метод точной квадратичной регуляризация для решения многоэкстремальных задач, который показал лучшие численные результаты в сравнении с генетическими и эволюционными методами, при решении многих тестовых задач [6].

Целью данной работы является сведение проблемы кластеризации данных к оптимизационной задаче и ее решение методом точной квадратичной регуляризации.

Постановка задачи и метод ее решения

Рассмотрим множество *m* точек $\{x^1,...,x^m\}$ в *n*-мерном евклидовом пространстве. Необходимо разбить это множество на *k* сферических кластеров таким образом, чтобы каждая точка попала только в один кластер и суммарное расстояние между центрами покрывающих точки шаров было максимальным. Будем покрывать множество заданных точек *n*-мерными шарами с радиусом *r*. Диаметр шара определяет максимально допустимое расстояние между точками одного кластера. Необходимо определить центры $\{z^1,...,z^k\}$ шаров $B_i = \{x \mid \| x - z^i \|^2 \le r^2\}, \quad i = 1,...,k$ так, чтобы $B_i \cap B_j = \Theta, \forall i \ne j$. Это условие равносильно системе неравенств

$$\| z^j - z^i \|^2 \ge 4r^2, i, j = 1,...k, i \ne j. \tag{1}$$

Каждая точка $x^i \in B_j$, что равносильно следующим ограничениям, при выполнении условий (1)

$$\prod_{i=1}^{k} (\| x^j - z^i \|^2 - r^2) \le 0, j = 1,...,m. \tag{2}$$

В качестве критерия оптимальности покрытия множества точек $\{x^1,...,x^m\}$ непересекающимися шарами, выберем максимизацию суммарного расстояния между центрами шаров. Это равносильно максимизации целевой функции

$$\sum_{j=1}^{k} \sum_{i=j+1}^{k} \| z^j - z^i \|.^2 \tag{3}$$

Решение задачи (1 – 3) определит центры шаров, которые разобьют множество точек $\{x^1,...,x^m\}$ на *k* кластеров. Данная постановка задачи кластеризации данных (1 – 3) проще существующих [6]. Задача (1 – 3) имеет *nk* искомых переменных $\{z^1,...,z^k\}$ и *m*+*k*(*k*–1)/2 ограничений (1), (2). Целевая функция (3) является выпуклой, а допустимое множество (1), (2) – невыпуклым, и задача (1 – 3) – многоэкстремальна. Классические методы ее решения, такие, например, как методы внутренней точки, позволяют

найти только локальное решение, при этом они не всегда могут найти допустимое решение. Поэтому используем метод точной квадратичной регуляризации [7] для решения задачи (1 – 3), который предназначен для решения многоэкстремальных задач.

Преобразуем задачу (1 – 3) к виду

$$\min\{z \mid -\sum_{j=1}^{k}\sum_{i=j+1}^{k}\| z^j - z^i \|^2 + s \le z, \prod_{i=1}^{k}(\| x^j - z^i \|^2 - r^2) \le 0, j = 1,...,m,$$
$$\| z^j - z^i \|^2 \ge 4r^2, i, j = 1,...,k\}, \tag{4}$$

где z – новая переменная, а параметр s удовлетворяет условию

$$s - \sum_{j=1}^{k}\sum_{i=j+1}^{k}\| z^{*j} - z^{*i} \|^2 \ge \| Z^* \|^2, Z^* = \{z^{*1},...,z^{*k}\},$$

Z^* – решение задачи (1 – 3). Используем замену $y = A\overline{Z}$, где матрица A порядка $(kn+1)\times(kn+1)$ равна

$$\begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ y_1 & y_2 & ... & y_{kn+1} \end{pmatrix},$$

а вектор $\overline{Z} = (Z, z)$, для преобразования задачи (4) к следующей

$$\min\{\| y \|^2 \mid -\sum_{j=1}^{k}\sum_{i=j+1}^{k}\| y^j - y^i \|^2 + s \le \| y \|^2, \prod_{i=1}^{k}(\| x^j - y^i \|^2 - r^2) \le 0,$$
$$j = 1,...,m, \| y^j - y^i \|^2 \ge 4r^2, i, j = 1,...,k\}. \tag{5}$$

Зададим такое значение параметра $q > 0$, при котором ограничения задачи

$$\min\{\| y \|^2 \mid -\sum_{j=1}^{k}\sum_{i=j+1}^{k}\| y^j - y^i \|^2 + s + (q-1)\| y \|^2 \le d, \prod_{i=1}^{k}(\| x^j - y^i \|^2 - r^2) +$$
$$q\| y \|^2 \le d, j = 1,...,m, q\| y \|^2 - \| y^j - y^i \|^2 + 4r^2 \le d, i, j = 1,...,k, q\| y \|^2 = d\} \tag{6}$$

будут выпуклыми, за исключением условия $q\| y \|^2 = d$. В задаче (6) значение новой переменной d необходимо определить. Пусть (y^0, d_0) – решение соответствующей выпуклой задачи

$$\min\{d \mid -\sum_{j=1}^{k}\sum_{i=j+1}^{k}\| y^j - y^i \|^2 + s + (q-1)\| y \|^2 \le d, \prod_{i=1}^{k}(\| x^j - y^i \|^2 - r^2) +$$
$$+ q\| y \|^2 \le d, j = 1,...,m, q\| y \|^2 - \| y^j - y^i \|^2 + 4r^2 \le d, i, j = 1,...,k, q\| y \|^2 \le d\}, \tag{7}$$

тогда, если условие $q\| y^0 \|^2 = d_0$ выполняется, то решение (y^0, d_0) – определяет оптимальное разбиение точек на k сферических кластеров. В противном случае, необходимо решать задачу

$$\max\{\| y \|^2 \mid -\sum_{j=1}^{k}\sum_{i=j+1}^{k}\| y^j - y^i \|^2 + s + (q-1)\| y \|^2 \le d, \prod_{i=1}^{k}(\| x^j - y^i \|^2 - r^2) +$$
$$+ q\| y \|^2 \le d, j = 1,...,m, q\| y \|^2 - \| y^j - y^i \|^2 + 4r^2 \le d, i, j = 1,...,k\} \tag{8}$$

и найти минимальное значение d^*, при котором выполняется условие $q\| y^* \|^2 = d^*$, где y^* – решение задачи (8) при фиксированном значении d^*. Будем решать задачу (8)

следующим образом. Выберем интервал $h > 0$ изменения переменной $d = d_0 + h$, где d_0 – решение задачи (7) и, решим для этого значения переменной d задачу (8) модифицированным методом внутренней точки. В этом методе на i-й итерации максимизация квадрата нормы вектора y на выпуклом допустимом множестве задачи (8), заменяется максимизацией линейной функции $(y^{i-1})^T y$, где y^{i-1} – решение задачи (8) на предыдущей итерации. Таким образом, на каждой итерации модифицированного метода решается задача выпуклой оптимизации. При увеличении переменной d значение целевой функции задачи (8) монотонно возрастает [7], пока не выполнится условие $q \| y^* \|^2 = d^*$, тогда точка y^* определит оптимальное разбиение точек на кластеры. Рассмотренная последовательность преобразований задачи (1 – 3) к эквивалентной задаче (8), при условии $q \| y \|^2 = d$, есть метод точной квадратичной регуляризации (EQR) [7].

Заметим, что увеличение параметра q не меняет решение задачи (8), но при таком увеличении происходит сглаживание локальных максимумов задачи (8). При больших значениях q каждое ограничение задачи (8)

$$\prod_{i=1}^{k} (\| x^j - y^i \|^2 - r^2) + q \| y \|^2 \le d, \ j = 1,...,m$$

будет стремиться к шару, а допустимая область этой задачи – к пересечению шаров.

Рассмотренная задача (1 – 3) решалась при фиксированном количестве кластеров. Метод EQR позволяет найти минимальное количество кластеров k^*. Для $k < k^*$ допустимое множество задачи (1 – 3) будет пустым. Это означает, что для решения задачи (8), при больших значениях d^*, будет выполняться условие $q \| y^* \|^2 < d^*$. Минимальное значение k, при котором допустимое множество задачи (8) не пусто, определит минимальное число кластеров.

Пример

На рис. 1 приведен расчет центров шаров кластеров для точек на плоскости

(1,2; 2,3; –1,3; 4,1; 4,–1; 3,–2; 5,0; 1,5; 4,4; 6,3; 5,5; 7,4; 5,–2)

при условии, что r^2 =3 и k =3. Задача (8) решалась при значениях параметров s = 200, q = 100. Центры шаров кластеров найдены в точках

(4.17082, –072361; 0.367281, 3.578124; 5.591608, 4.683216).

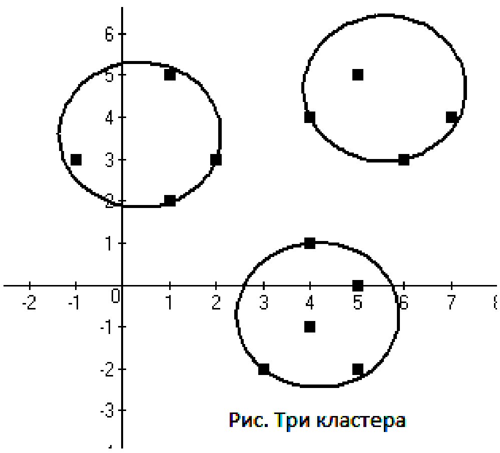


Рисунок 1 – Три кластера

Выводы

В работе приведена новая оптимизационная постановка задачи построения сферических кластеров в n-мерном евклидовом пространстве. Для решения полученной многоэкстремальной задачи использовался метод точной квадратичной регуляризации, эффективность которого проверена в многочисленных экспериментах.

Литература

1. Хант Э. Искусственный интеллект / Э. Хант; пер с англ. Д.А. Белова и Ю.И. Крюкова ; под. ред. В.Л. Стефанюка. – М. : Мир, 1978. – 560 с.
2. Ту Дж. Принципы распознавания образов /Дж. Ту, Р. Гонсалес ; пер. с англ. И.Б. Гуревича ; под. ред. Ю.И. Журавлева. – М. : Мир, 1978. – 413 с.
3. Мандель И.Д. Кластерный анализ / И.Д. Мандель. – М. : Финансы и статистика. – 1988. – 176 с.
4. Kenneth V.P. Differential Evolution. A Practical Approach to Global Optimization / V.P. Kenneth, R.M. Storn, J.A. Lampinen. – Berlin Heidelberg: Springer-Verlag, 2005. – 542 p.
5. Сокуренко В.М. Числове дослідження стохастичних методів безперервної глобальної оптимізації / В.М. Сокуренко, В.С. Неділюк // Наукові вісті НТУУ «КПІ». – 2012. – № 1. – С. 81-87.
6. Sherali H.D. A Global Optimization RLT-based Approach for Solving the Hard Clustering Problem / H.D. Sherali, D. Jitamitra // Journal of Global Optimization. – 2005. – 32. – P. 281-306.
7. Косолап А.И. Метод квадратичной регуляризации для решения систем нелинейных уравнений / А.И. Косолап // Журнал обчислювальної та прикладної математики. – 2010. – №4. – С. 44-50.

Literatura

1. Hunt E.B. Artificial Intelligence. Academic Press. Nev York, San Francisco, London, 1975. 468 p.
2. Tou J.T., Gonzalez R.C. Pattern Recognition Principles. Addison-Wesley Publishing Company. London-Amsterdam-Dom Mills, Ontario-Sydney-Tokyo. 1974. 378 p.
3. Mandel I.D. Cluster Analisys. Finances and Statistica. Moscow. 1988. 176 p. (rus)
4. Kenneth V.P., Storn R.M., Lampinen J.A. Differential Evolution. A Practical Approach to Global Optimization. Springer-Verlag. Berlin Heidelberg. 2005. 542 p.
5. Sokurenko V.M. Naukovi visti NTUU "KPI". No. 1. 2012. Pp. 81–87. (rus)
6. Sherali H.D., Jitamitra D. J. Global Optim. No. 32. 2005. Pp. 281–306.
7. Kosolap A.I. J. Comp. & Appl. Math. No. 4. 2010. Pp. 44–50. (rus)

A.I. Kosolap

Method of an Exact Quadratic Regularization Into Clustering Problem of Data

Partition of data into clusters is the most important problem of an artificial intellect and the theory of pattern classification. Data is represented by points of n-dimensional space. We consider this problem in n-dimensional space and use the spherical clusters. It is necessary to find the centers of the spheres which contain all points. Spheres should not intersect and the sum of the distances between their centers should be maximum. Such a statement concerning this problem for clustering the data is new and more simple [6].

This type of problem clustering of data is transformed to the maximization of the vector norm on nonconvex set. This problem is multiextreme. We use the method of exact quadratic regularization for its solution. This method transforms a multiextreme problem to maximization of a norm vector on a convex set. This convex set is approximated by the intersection of spheres. We use a dual method for the solution of this problem. This auxiliary problem is used for searching the starting point of the method of exact quadratic regularization. We use the modification of an interior point method for the solution of the problem of a maximum vector norm on convex set.

The method of the exact quadratic regularization has shown its superiority over the genetic and evolutionary methods at the solution of numerous test problems. The given method can be used for the data on clusters that was confirmed by numerous experiments.

Статья поступила в редакцию 19.11.2012.